ADDED VARIABLE PLOTS IN LINEAR REGRESSION

Bradford W. Johnson and Robert E. McCulloch
Dept. of Applied Statistics Graduate School of Business
University of Minnesota University of Chicago

University of Minnesota
School of Statistics
Technical Report No. 475
June 1986

This article discusses three well known graphical methods for identifying the need for and nature of transformations of explanatory variables in linear regression. Analysis and example show these methods can be very misleading even when the random additive error is negligible. New methods are proposed and examples of their use are given.

Key Words: Regression diagnostics, added variable plot, residual plus component plot, piecewise linear approximation

1. INTRODUCTION

In the standard linear model $Y_{n\times 1} = X_{n\times k}\,\beta_{k\times 1} + \epsilon_{n\times 1}$, with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I$, summary statistics such as parameter estimates and R^2 values are unable to detect the failure of the model to fit the data. This has led to the development of diagnostic tests and plots designed to identify important departures from the model such as the presence of outliers, heterogeneity of variances, and nonlinearity. The recent books by Cook and Weisberg(1982) and Belsley, Kuh and Welsch(1980) are surveys of diagnostic methods in regression.

Often in formulating the above linear model the researcher feels strongly that an explanatory variable included in the matrix X influences the response Y but is not really sure that it is not some function of the explanatory variable that is linearly related to the mean of the response rather than the variable as he happens to have measured it. In this paper we discuss graphical methods for determining a function f such that $Y = X\beta + \gamma f(Z) + \epsilon$ is a better model then that obtained by simply assuming that f is the identity map $f(x)=x$, where Z is a known explanatory variable and the vector $f(Z) = (f(Z_1), f(Z_2), \ldots, f(Z_n))$. In this exploratory process we feel that graphical methods are more appropriate than any attempt at more classical probabilistic inference involving the assumption of a parametric family of functions f. Graphical methods allow the user to see much of the data at once rather than just a summary so that

anomalies and patterns may be uncovered. However, it must be borne in mind that the graphical methods discussed in this paper are meant to be suggestive rather than conclusive. We will discuss the behavior of commonly used methods and propose a new one.

The new method is based on the assumption that the function f is sufficiently smooth for a simple linear approximation to f to work well locally. We must make some kind of restriction on the function f. If any function is allowed it is usually possible to find an f such that $Y = f(Z)$, a useless result. Thus we cluster the observations according to the value of the variable Z. Within a cluster, where Z does not vary much, we assume f is linear and allow the slope and intercept to vary from cluster to cluster. We feel that the use of clusters is an important tool in exploratory analyses of this kind. For another example of its use see Landwehr et al (1984).

Perhaps the most commonly used methods are:

(i) The simple residual plot: plot the vector of O.L.S. residuals for the regression of Y on X vs the vector Z; $(Y-X\hat{\beta})$ vs Z. Here $\hat{\beta}$ minimizes $(Y-X\beta)^T(Y-X\beta)$.

(ii) The added variable plot: plot the vector of residuals of Y on X vs the vector of residuals of Z on X; $(Y-X\hat{\beta})$ vs $(Z-X\hat{\alpha})$. Here $\hat{\beta}$ minimizes $(Y-X\beta)^T(Y-X\beta)$ and $\hat{\alpha}$ minimizes $(Z-X\alpha)^T(Z-X\alpha)$.

(iii) The residual plus component plot: plot $(Y-X\hat{\beta}-Z\hat{\gamma}) + Z\hat{\gamma}$ vs Z where $(\hat{\beta},\hat{\gamma})$ minimizes $(Y-X\beta-\gamma Z)^T(Y-X\beta-\gamma Z)$.

(iv) Introduce a new parameter λ and let $Y = X\beta + \gamma Z^{(\lambda)} + \epsilon$ where $Z^{(\lambda)}$ is defined in Cook and Weisberg (1982) page 60.

Method (iv) was introduced by Box and Tidwell (1962). The other three methods are discussed in Cook and Weisberg (1982) and Mallows (1982).

A new method, the ACE algorithm (Breiman and Friedman (1985)), appears to be the ultimate tool for determining transformations in linear models. The three graphical methods above and the one presented below are, however, simple to use and understand in the context of standard regression packages.

Since the emphasis of this paper is on informal graphical techniques, method (iv) will not be discussed. Section 2 discusses the simple residual plot, the added variable plot and the residual plus component plot. Section 3 introduces a new alternate method. In section 4, we conclude the paper with recommendations and suggestions for future work.

2. STANDARD METHODS

Given a matrix X, $P_{(X)}$ shall denote the orthogonal projection onto the linear subspace generated by the columns of X. $Q_{(X)}$ shall denote the orthogonal projection onto the subspace of vectors orthogonal to the columns of X. Thus $Q_{(X)} = I - P_{(X)}$ where I is the identity matrix. When X has full rank it is well known that $P_{(X)} = X(X^TX)^{-1}X^T$.

The solution to the problem of minimizing the sum of squares $(Y-X\beta)^T(Y-X\beta)$ with respect to β, is obtained by choosing a $\hat{\beta}$ such that $X\hat{\beta}$ is the orthogonal projection of Y onto the column space of X and $Y-X\hat{\beta}$, the residuals, is the orthogonal projection of Y onto the subspace of vectors orthogonal to the columns of X. Thus, the vector of residuals equals $Q_{(X)}Y$. This geometrical viewpoint has proven to be of use in thinking about regression problems. In this section we explore the three graphical methods in section 1 from a geometrical viewpoint. In order to do this we rewrite the plotted vectors in terms of projection matrices P and Q for the relevent subspaces.

Clearly we have:

(i) the simple residual plot: $Q_{(X)}Y$ vs Z

(ii) the added variable plot: $Q_{(X)}Y$ vs $Q_{(X)}Z$.

Expressing the residual plus component plot in terms of projections is somewhat more difficult. The development of the following expression is left to the appendix.

(iii) the residual plus component plot: ΨY vs Z, where

$$\Psi = \left(I + \frac{Z^T Z}{Z^T Q_{(X)} Z} P_{(X)} P_{(Z)} \right) Q_{(X)}$$

Note that if Z is perpendicular to the column space of X then $Q_{(X)}Z = Z$ and $P_{(X)} P_{(Z)} = 0$, so that $\Psi = Q_{(X)}$ and all of the three plots are the same.

Although Ψ looks complicated, it is easily understood by breaking R^n up into subspaces and seeing how Ψ acts within each particular subspace. Any vector S in R^n may be written as $S = X\alpha + cZ + V$ where α is a $k \times 1$ vector, c is a scalar, and V is a vector orthogonal to both Z and the columns of X. We have decomposed R^n into three subspaces: the column space of X, the one dimensional subspace along Z, and the subspace orthogonal to both Z and the columns of X.

It is easy to show the following three properties of Ψ:

(i) $\Psi(Z) = Z$ (2.1)

(ii) $\Psi(U) = 0$ for all vectors U in the column space of X. (2.2)
(i.e. $\Psi(X\alpha) = 0$ for all α in R^k.)

(iii) $\Psi(V) = V$ for all V orthogonal to both Z and the columns of X. (2.3)

We then have $\Psi(S) = \Psi(X\alpha + cZ + V) = cZ + V$. Note that $\Psi^2 = \Psi$. Writing vectors in the decomposed form will make the geometry of the three plots clear, since both the actions of Ψ and $Q_{(X)}$ may then be expressed simply.

Given our model $Y = X\beta + \gamma f(Z) + \epsilon$, all three plots consist of plotting a vector obtained by applying a linear map to the random Y against the constant vector Z or $Q_{(X)}Z$. Rather than the random Y we shall consider the expectation of Y. Since the maps applied to Y are linear this will correspond to studying the average behaviour of the plots or the behaviour of the plots when the error is relatively small. We then have the following three plots:

(i) srp: $Q_{(X)}E(Y)$ vs Z

(ii) avp: $Q_{(X)}E(Y)$ vs $Q_{(X)}Z$

(iii) rpc: $\Psi E(Y)$ vs Z

where srp, avp, and rpc correspond to simple residual plot, added variable plot, and residual plus component plot respectively.

The expectation of Y, given the assumption $E(\epsilon) = 0$, is $X\beta + f(Z)$ where, without loss of generality, we have let $\gamma = 1$. Since $f(Z)$ is a vector in R^n we may write $f(Z) = X\alpha + cZ + V$, where, as above, α is a vector in R^k, c is a scalar, and V is a vector orthogonal to both Z and the columns of X.

Then $E(Y) = X\beta + f(Z)$

$= X\beta + (X\alpha + cZ + V)$

$= X(\beta+\alpha) + cZ + V.$

So, $Q_{(X)}E(Y) = cQ_{(X)}Z + V$, and

$\Psi E(Y) = cZ + V$

where we have used the properties (2.1), (2.2), and (2.3) of Ψ and the obvious facts, $Q_{(X)}X = 0$ and $Q_{(X)}V = V$.

Using the above, we may now write the three plots under study as:

(i) srp: $cQ_{(X)}Z + V$ vs Z (2.4)

(ii) avp: $cQ_{(X)}Z + V$ vs $Q_{(X)}Z$ (2.5)

(iii) rpc: $cZ + V$ vs Z. (2.6)

Let us now examine the three plots assuming that the function f is linear. If f is linear then α and V are 0 so that $f(Z) = X\alpha + cZ + V = cZ$. The three plots become:

(i) srp (f linear): $cQ_{(X)}Z$ vs Z (2.7)

(ii) avp (f linear): $cQ_{(X)}Z$ vs $Q_{(X)}Z$ (2.8)

(iii) rpc (f linear): cZ vs Z. (2.9)

We see immediately that the simple residual plot fails in a basic way. Even if f is linear and the error is negligible, we do not get a linear plot unless Z is orthogonal to the columns of X in which case, as noted above, all of the three plots are indentical. The other two plots may be viewed as efforts to correct this problem. In going from the simple residual plot to the added variable plot (equation (2.4) to (2.5)) we go from plotting against Z to plotting against $Q_{(X)}Z$. This adjustment results in a linear plot when f is linear (equation 2.8). In going from the simple residual plot to the residual plus component plot (equation (2.4) to (2.6)) we still plot against Z, but add $cP_{(X)}Z$ to the vector being plotted. For f linear, this results in the linear plot

cZ vs Z given in equation (2.9).

In an application of these methods to data we do not know whether f is linear not. The question then is, if the plots are linear do we have information indicating that f is linear. The answer is no. From equations (2.5) and (2.6) we see that the plots are linear if and only if V = 0. So a linear plot indicates that f(Z) is of the form $X\alpha + cZ$ which need not be linear.

The problem is that in all three plots $X\alpha$, the part of f in the column space of X, is lost. The residual plus component plot adds on $cP_{(X)}Z$, what would have been lost if f were linear. If, however, f is not linear, adding on $cP_{(X)}$ is an arbitrary act which may easily do more harm than good in trying to get a picture of f. Similarly, plotting against $Q_{(X)}Z$ rather than Z makes the added variable plot linear when - is, but may cause undesireable distortion in general.

2.1 An Example

We now construct a simple example to illustrate the ideas of the preceeding paragraph. We will construct the example so than the simple residual plot gives an almost exact reproduction of the function f, while the adjustments made to produce the other two plots result in misleading plots. Also, there will be no error in the example so that $Y = E(Y) = X\alpha + f(Z)$.

We use the notation of this section to describe our example so

that $f = X\alpha + cZ + V$. For simplicity we will let X consist of only one column and use $\alpha = c = 1$ so that $f = X + Z + V$, where as above, the vector V is orthogonal to both X and Z. From equation (2.4) we see than in order for the simple residual plot to be a plot of f(Z) vs Z we need to have $Q_{(X)}Z = Z + X$. We will choose Z and X so that $Z = (Z+X) + (-X)$ is an orthogonal decomposition of Z, in which case $P_{(X)}Z = -X$ and $Q_{(X)}Z = Z + X$ as desired.

We let $Z^T = (-30,-29,\ldots-1,0,1,\ldots29,30)$.

Let $\tau = .3Z + .7(Z^3/600)$. Let $X = -Q_{(\tau)}Z$. We then have, $Z = P_{(\tau)}Z + Q_{(\tau)}Z$, with $Q_{(\tau)}Z = -X$ and $P_{(\tau)}Z = Z + X$ so than $Z = (Z+X) + (-X)$ is an orthogonal decomposition. Finally let $V = (Z^2)/30$. Note that X is a linear combination of Z and Z^3. Since Z is symmetric about 0 we see than V is orthogonal to both Z and X because V is an even function and both Z and Z^3 are odd.

So, with Z, X, and V, as just defined above and $f = X + Z + V$, and $Q_{(X)}Z = Z + X$, we see from equations (2.4), (2.5), and (2.6) that our three plots are:

(i) srp: $Z + X + V = f(Z)$ vs Z

(ii) avp: $f(Z)$ vs $Z + X$

(iii) rpc: $Z + V$ vs Z .

Note that $Z + V$ is a quadratic function of Z while $f(Z)$ is a

cubic function of Z so that the residual plus component plot is quite misleading. The added variable plot plots f(Z) versus a cubic function of Z causing considerable distortion in the plot.

One final adjustment is needed to make the example meaningful. The expectation of Y in this example is X + f(Z) = X + (Z + X + V) = 2X + (Z + V) = g(Z) so that in the model $E(Y) = X\alpha + f(Z)$, where f is a reasonably smooth function, there is clearly an identification problem. We correct this by using X' instead of X where $X' = X + E$ and $E = (e_1, e_2, \ldots e_{61})$ with the e_i i.i.d N(0,1). Now X' is not a smooth function of Z so that the identification problem goes away.

With X', Z, and f(Z) as defined above we let $Y = X' + f(Z)$. Figure 2.1 is a plot of f(Z) vs Z. Now forget about the X we used above to construct the example and let X be the matrix whose first column is a column of ones and whose second column is X'. Figure 2.2 is a plot of $Q_{(X)}Y$ vs Z (the simple residual plot). This plot clearly indicates the correct cubic form of f. Figure 2.3 is a plot of $Q_{(X)}Y$ vs $Q_{(X)}Z$ (the added variable plot). This plot is significantly distorted. Figure 2.3 is a plot of ΨY vs Z (the residual plus component plot). This plot incorrectly indicates that f is a quadratic function of Z. Note that the horizontal scale for figure 2.3 is not the same as for the others.

3. AN ALTERNATIVE PLOT

In suggesting an alternative plot we proceed without anticipating any particular functional form for f(z). In particular, the possible linearity of f(z) in no way motivates the procedure.

We wish to consider as candidates for f(z) as wide a class of functions as possible that is still restrictive enough to avoid overfitting. We feel that a reasonable choice for this is the class of piecewise linear functions.

In order to estimate f near z_0 under these assumptions, we need a set of observations z_j, in an interval around z_0. For z_j in the i^{th} interval we thus model

$$f(z_j) \approx a_i + b_i(z_j - z_0).$$

In other intervals the model for f(z) has the same form although the constants a_i, b_i generally will be different.

Suppose we have the data y_i, x_i, and z_i, where y_i and z_i are scalers and x_i is p dimensional where, as above, the y's are the response variables, the x's are vectors of explanatory variables, and the z's are explanatory variables that may require transformation by the unknown function f. In order to implement our piecewise linear representation of f we first cluster our observations by their z value. The set of n observations is chopped

up into disjoint subsets, or clusters, so that within each subset the values of the variable z do not vary much relative to the overall variation in z. It is also important to choose the subsets so that the number of points in a subset does not vary too greatly from subset to subset so that our available information, the data, is not used up pinning down the function f at a few points while gaining little information about it's overall behaviour. A third consideration is the number of parameters we are adding to the model. For each cluster we must estimate a_i and b_i. If we have too many clusters, with each cluster having only a few observations, we may end up with too many parameters for the amount of data we have.

Obviously it may be very difficult to balance all of these factors. Rather than attempting an analytic determination of the "optimal" choice (a difficult task), we suggest an exploratory approach in which various reasonable partitions are entertained and the sensitivity of the outcome to the choice is examined. If the outcome is insensitive to changes in those model assumptions about which we are uncertain (such as the correct partition scheme) then we are reassured. As we shall see below, once the partition is chosen it is easy to implement the procedure so that there is no difficulty in trying various partition schemes. In comparing the method of this paper to a technique such as ACE (Breiman and Friedman(1985)), we see that the user must directly confront the nature of the information at hand rather than let a program handle everything. The additional effort required to choose the partition

scheme may well be worth the added awareness gained by the user.

Once the partition scheme has been chosen we relabel the observations so that the ij^{th} observation is the j^{th} observation in the i^{th} partition. We then have,

$$y_{ij} = x_{ij}\beta + a_i + b_i (z_{ij} - \bar{z}_i) + \varepsilon_{ij}, \qquad (3.1)$$

where we have used the appoximation,

$$f(z_{ij}) \approx a_i + b_i (z_{ij} - \bar{z}_i),$$

and $\bar{z}_i$ is the mean of the z values of the observations in the i^{th} partition. The above model is then a linear model and is easily fit using the standard packages.

An even simpler approach assumes that f may be approximated by a constant within a partition, so that b_i is zero. Our model is then,

$$y_{ij} = x_{ij}\beta + a_i + \varepsilon_{ij}. \qquad (3.2)$$

Again we have a linear model which is fit using standard packages.

Once the estimates have been computed we plot the estimates of the a_i's against the $\bar{z}_i$'s. This plot then is examined in the hope

the a natural function f is evident. We emphasize that the procedure is exploratory and suggestive rather then conclusive. To some extent the uncertainty of the plot is captured by the standard errors of the estimates of the a_i's. Standard multiple comparison techniques may be used to obtain intervals for each estimate a_i which reflect the overall uncertainty in the plot. These intervals could then be incorporated into the plot. However we prefer to think of the method as exploratory and just plot the estimates. If a function is easily inferred from the plot, the user must then incorporate an analytic representation of the function and make judgements about whether or not the new model is an improvement over the simple linear model.

We also considered a third method which involves the repeated use of observations. Again the observations are grouped by similar z values, but now we no longer require the groups to be disjoint. In the examples we studied we did not find this method superior to the other methods and it is much more difficult to use than the other methods.

To compute the estimates of the a_i's we wrote a short fortran program which first sorts the data according to the z values and then, given a partition scheme, generates matrices which correspond to the regresson problems indicated in (3.1) or (3.2) above. For the model (3.2) the matrix would consist of the p columns corresponding to the explanatory variables in the x's, and indicator columns for each partition, where the indicator column for the i^{th}

partition has entry 0 for observations in the other partitions and a 1 for observations in the i^{th} partition. The vector of y's is then regressed on the matrix with no intercept in the model and the estimates of the a_i's are the estimates corresponding to the indicator variables. For the model (3.1) the matrix would consist of all the columns used for model (3.2) with an additional column for each partition where the column for the i^{th} partition has entry 0 for the observations in the other partitions and a $(z_{ij} - \bar{z}_i)$ value for the j^{th} observation in the i^{th} partition. Again the estimates of the a_i's are the estimates corresponding to the indicator variables. We now present some examples.

3.1 Example 1

We apply the methods introduced above to the tree data from the Minitab Student Handbook (Ryan, Joiner, and Ryan(1976) page 278). The data consists of measurements on the volume, height, and diameter at 4.5 feet above ground level for a sample of 31 black cherry trees in the Allegheny Forest, Pennsylvania. The goal of the study is to be able to infer the volume of a tree from it's height and diameter. We will consider a regression of volume on height and diameter and look for a function of diameter which improves the model.

We must decide on reasonable partitions of the observations based on the diameter values. Figure 3.1.1 is a plot of the

diameter versus the case index. This plot is useful in choosing the partitions since it is easy to pick out clusters by eye. The following table gives the data and two partition schemes. The number under the partition headings indicates which partition the observation belongs to. The symbol – indicates that the observation has been deleted. An observation is deleted when there are not enough observations having sufficiently similar diameter values for estimation of f.

CASE	HEIGHT	DIAMETER	VOLUME	PARTITION1	PARTITION2
1	70	8.3	10.3	1	1
2	65	8.6	10.3	1	1
3	63	8.8	10.2	1	1
4	72	10.5	16.4	2	2
5	81	10.7	18.8	2	2
6	83	10.8	19.7	2	2
7	66	11.0	15.6	2	2
8	75	11.0	18.2	2	3
9	80	11.1	22.6	2	3
10	75	11.2	19.9	3	3
11	79	11.3	24.2	3	3
12	76	11.4	21.0	3	4
13	76	11.4	21.4	3	4
14	69	11.7	21.3	3	4
15	75	12.0	19.1	3	4
16	74	12.9	22.2	4	5
17	85	12.9	33.8	4	5
18	86	13.3	27.4	4	5
19	71	13.7	25.7	4	5
20	64	13.8	24.9	5	6
21	78	14.0	34.5	5	6
22	80	14.2	31.7	5	6
23	74	14.5	36.3	5	6
24	72	16.0	38.3	–	7
25	77	16.3	42.6	–	7
26	81	17.3	55.4	6	7
27	82	17.5	55.7	6	8

28	80	17.9	58.3	6	8
29	80	18.0	51.5	6	8
30	80	18.0	51.0	6	8
31	87	20.6	77.0	–	–

Figure 3.1.2 is the plot resulting from using the simple approximation of model (3.2) above and the first partition scheme. Figure 3.1.3 results from the linear approximation (3.1) above and the first partition scheme. Figures 3.1.4 and 3.1.5 are models (3.2) and (3.1) repectively with the second partition scheme. All of the plots suggest that a cubic or quadratic f applied to diameter might improve the model. Using the data plotted in figure 3.1.4 we regress the estimated f values on a constant term and the square of the z values corresponding to the f values. The resulting R^2 is .98 which suggests that the simple function $f(z) = z^2$ might be useful.

Regressing volume on diameter and height we obtain on R^2 of .95. Regressing volume on diameter squared and height we obtain an R^2 of .97 which is an improvement, although slight. Consider, however, the residual plots in the two regressions. Figure 3.1.6 is a plot of the residuals from the regression of volume on diameter and height versus diameter. The points in the plot appear to fall and then rise again suggesting that there is something wrong with the model. Figure 3.1.7 is the plot of the residuals from the regression of volume on diameter squared and height versus diameter squared. This residual plot looks much better.

The above is not a complete analysis of the tree data, but does show how the proposed methods can lead to better models.

3.2 Example 2

Our second example is taken from section 2.1 of Weisberg(1983). We are regressing fuel on tax and dlic where fuel is per capita motor fuel consumption in gallons, tax is the tax on fuel in cents per gallon, and dlic is the proportion of the population with drivers licenses. We shall apply the methods to the explanatory variable tax. The following table gives the data and partition scheme.

CASE	FUEL	DLIC	TAX	PARTITION
1	640	56.6	5.0	-
2	782	67.2	6.0	-
3	644	69.2	6.58	-
4	632	60.3	7.0	1
5	865	72.4	7.0	1
6	714	54.0	7.0	1
7	603	57.2	7.0	1
8	649	66.3	7.0	1
9	699	56.3	7.0	1
10	587	62.6	7.0	1
11	554	51.3	7.0	1
12	571	51.8	7.0	1
13	704	58.6	7.0	1
14	968	67.2	7.0	1
15	635	58.6	7.0	1
16	524	59.3	7.0	1
17	566	60.8	7.0	1
18	591	50.8	7.0	1
19	610	62.3	7.0	1
20	498	55.2	7.0	1
21	525	57.4	7.0	1
22	508	54.5	7.0	1
23	631	57.9	7.5	2
24	414	52.9	7.5	2

25	628	54.7	7.5	2
26	471	52.5	7.5	2
27	410	54.4	8.0	3
28	487	48.7	8.0	3
29	577	57.8	8.0	3
30	580	53.0	8.0	3
31	574	56.3	8.0	3
32	540	60.2	8.0	3
33	464	52.9	8.0	3
34	344	45.1	8.0	3
35	467	55.3	8.0	3
36	577	54.8	8.0	3
37	648	66.3	8.5	4
38	460	55.1	8.5	4
39	640	67.7	8.5	4
40	524	57.2	9.0	5
41	561	58.0	9.0	5
42	510	57.1	9.0	5
43	534	49.3	9.0	5
44	464	51.1	9.0	5
45	566	54.4	9.0	5
46	547	51.7	9.0	5
47	541	52.5	9.0	5
48	457	57.1	10.0	–

The rational behind the partition scheme is obvious. Figure 3.2.1 is the plot obtained from the simple approximation (3.2). Clearly there is no need to consider the linear approximation of (3.1) in this example. Figure 3.2.1 indicates that the linear model works well for taxes up to 8.5, but not beyond.

3.3 Example 3

Our third example is data set 25 from the book Graphical Methods for Data Analysis by Chambers, Cleveland, Kleiner, and Tukey(1983). Thirty rubber specimens were rubbed with an abrasive material. The variables are, hardness in degrees Shore, tensile strength in kilograms per square centimeter, and abrasion loss (the amount of material rubbed off) in grams per horsepower-hour . We will use the abreviations hard, ten, and loss. We consider a regression of loss on ten and hard. We shall apply the methods the variable ten.

Figure 3.3.1 is a plot of ten versus the case index. Again, we use this plot to pick out reasonable partition schemes. The following table give the data and two partition schemes.

CASE	LOSS	HARD	TEN	PARTITION1	PARTITION2
1	64	88	119	1	1
2	148	86	127	1	1
3	114	89	128	1	1
4	215	81	134	1	2
5	267	74	144	1	2
6	340	59	146	2	2
7	283	65	148	2	3
8	155	82	151	2	3
9	219	71	151	2	3
10	249	59	161	2	4
11	341	51	161	3	4
12	97	83	161	3	4
13	372	45	162	3	5
14	186	80	165	3	5
15	196	68	173	3	5
16	32	81	180	4	6

17	128	75	188	4	6
18	166	60	189	4	6
19	82	79	196	4	7
20	228	56	200	4	7
21	221	53	203	5	7
22	164	64	210	5	8
23	113	68	210	5	8
24	45	86	219	5	8
25	55	81	224	5	9
26	154	66	231	6	9
27	136	71	231	6	9
28	175	61	232	6	10
29	206	55	233	6	10
30	112	71	237	6	10

The first partition scheme was used for figure 3.3.2 and figure 3.3.3, which correspond to the models (3.2) and (3.1) above, respectively. Note that the two plots appear quite different. Figure 3.3.4 is obtained from the second partition scheme and the simpler model (3.2). Figures 3.3.3. and 3.3.4 are closer to agreement. In this example no simple function f is suggested by the plots. The sensitivity of the plots to the partition scheme and choice of model, as well as the irregular appearance of the plots, suggests that the data set should be carefully studied before the simple linear model is adopted.

3.4 Example 4

For this example we have constructed a dataset such that the loss of $P_{(X)}f$ causes the plots discussed in section 2 to give misleading results. With the right choice of group size, the new methods clearly indicate the correct form of the function. We chose X to be orthogonal to Z, which will make all of the plots from section 2 identical.

In this example, Z consists again of the integers from -30 to 30, and $f(z) = (z/6)^2$. X is shown in figure 3.4.2 as a function of Z. Note that the function relating X and Z is neither linear nor continuous, and that f(Z) and X coincide for the larger absolute values of z. $Y = X + f(Z) + \epsilon$, where ϵ is N(0,3). Figure 3.4.1 is a plot of f(Z) vs. Z. Figure 3.4.3 is a plot of $f(Z) + \epsilon$ vs. Z, a plot which displays the information available if β were known. Figure 3.4.4 is a plot of $Q_{(X)}Y$ vs. Z, the methods of section 2. Figures 3.4.5 and 3.4.6 are the piecewise linear method with partition sizes 7 and 3 respectively.

Figure 3.4.4 shows that none of the methods from section 2 work at all for this example. In figure 3.4.6 we show the piecewise linear method (3.1), with group size 3, which is almost exact.

Although the picture changed radically when we changed the partition size and applied the linear approximation we are reassured by noting that the striking figure 3.4.6 is obtained with smaller partition sizes which correspond to less restricitive assumptions on

the function f. The danger in small partitions is that the procedure might start chasing the random error rather than the function f, in which case the plot appear less smooth. This is certainly not the case in figure 3.4.6. This example illustrates the importance of trying different partition schemes.

3.5 Example 5

In this example we apply the method introduced in this section to the constructed data set of section 2. Recall that the example was constructed by making everything in sight a smooth function of Z and then fuzzing the X vector by adding noise to it. This example is then an interesting test for the method. If it can pick out the part of Y which is a smooth function of Z it won't be fooled by X which is close to being a smooth function of Z. Recall also, that in this example there is no error in the Y vector so that $Y = X\beta + f(Z)$.

We try two different partition schemes. For each partition scheme we use the linear approximation method. In figure 3.5.1 we see the plot that results from using 10 partitions. The nature of the Z vector makes the appropriate partitioning scheme obvious (recall Z = (-30,-29,...29,30)). The plot clearly indicates the correct form of the function f. In figure 3.5.2 we see the plot that results from using 20 partitions. Again the correct form of f is apparent.

4. CONCLUSION

The piecewise linear method seems to work well and is easy to use within standard regression packages. The new methods have been specifically designed to indicate a transformation assuming only that the function is reasonably smooth. The other methods discussed seem to have been constructed with the assumption that f is roughly linear.

Note that the new methods can easily be extended to other models such as generalized linear models.

ACKNOWLEDGEMENT

The authors thank Professor R. D. Cook for suggesting the topic and for many helpful discussions. We also take this opportunity to thank him for his excellent teaching.

This work was funded by the Social Sciences and Humanities Research Council of Canada and NIH grant NIGMS 25271.

APPENDIX

We now show that the residual plus component plot is of the form ΨY vs Z where Ψ is given in section 2.

Let the matrix $M = (X,Z)$ and the vector $U = Q_{(X)}Z$. We need the following lemma.

Lemma: $Q_{(M)} = Q_{(U)}Q_{(X)}$, assuming M has full rank.

proof:

We first calculate $P_{(M)} = M(M^TM)^{-1}M^T$.

$$M^TM = \begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ \end{bmatrix}.$$

By a standard result on the inverse of partitioned matrices (see for example Rao(1973), chapter 1), $(M^TM)^{-1} =$

$$\begin{bmatrix} (X^TX)^{-1} + \dfrac{(X^TX)^{-1}X^TZZ^TX(X^TX)^{-1}}{Z^TQ_{(X)}Z} & \dfrac{-(X^TX)^{-1}X^TZ}{Z^TQ_{(X)}Z} \\ \dfrac{-Z^TX(X^TX)^{-1}}{Z^TQ_{(X)}Z} & \dfrac{1}{Z^TQ_{(X)}Z} \end{bmatrix}$$

Straightforward multiplication, substitution of $I - Q_{(X)}$ for $P_{(X)}$, and simplification then gives,

$$M(M^TM)^{-1}M^T = I - Q_{(X)} + \frac{Q_{(X)}ZZ^TQ_{(X)}}{Z^TQ_{(X)}Z}.$$

$$\text{So, } Q_{(M)} = I - P_{(M)} = Q_{(X)} - \frac{Q_{(X)}ZZ^TQ_{(X)}}{Z^TQ_{(X)}Z}$$

$$= (I - \frac{Q_{(X)}ZZ^TQ_{(X)}}{Z^TQ_{(X)}Z})Q_{(X)}$$

$$= (I - \frac{UU^T}{U^TU})Q_{(X)}$$

$$= Q_{(U)}Q_{(X)}.$$

Where we have used $Q^2_{(X)} = Q_{(X)}$. □

Result: Suppose $\hat{\beta}$ and $\hat{\gamma}$ are the least squares estimates of β and γ in the model $Y = X\beta + \gamma Z + \epsilon$. Then,

$$(Y - X\hat{\beta} - \hat{\gamma}Z) + \hat{\gamma}Z = \Psi Y,$$

where Ψ is defined in section 2.

proof:

First note than $(Y - X\hat{\beta} - \hat{\gamma}Z) = Q_{(M)}Y$, for $M = (X,Z)$.

Also, $\hat{\gamma} = \dfrac{Z^T Q_{(X)} Y}{Z^T Q_{(X)} Z}$.

So,

$$Q_{(M)}Y + Z\hat{\gamma} = Q_{(M)}Y + \frac{Z\, Z^T Q_{(X)} Y}{Z^T Q_{(X)} Z}$$

$$= \left(Q_{(M)} + \frac{Z\, Z^T Q_{(X)}}{Z^T Q_{(X)} Z}\right) Y$$

(Now note that by the lemma we may write $Q_{(M)} = Q_{(U)} Q_{(X)}$)

$$= \left(Q_{(U)} + \frac{Z\, Z^T}{Z^T Q_{(X)} Z} \right) Q_{(X)} Y$$

$$= \left(I - \frac{Q_{(X)} Z\, Z^T Q_{(X)}}{Z^T Q_{(X)} Z} + \frac{Z\, Z^T}{Z^T Q_{(X)} Z} \right) Q_{(X)} Y$$

$$= \left(I - \frac{Q_{(X)} Z\, Z^T - Z\, Z^T}{Z^T Q_{(X)} Z} \right) Q_{(X)} Y$$

$$= \left(I + \frac{Z^T Z}{Z^T Q_{(X)} Z} P_{(X)} P_{(Z)} \right) Q_{(X)} Y$$

$= \Psi Y$ as desired. □

REFERENCES

BELSLEY, D. A., KUH, E., and WELSCH, R. (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley.

BOX, G. E. P., and TIDWELL, P. W. (1962), Transformations of the Independent Variables, Technometrics, 4, 531-550.

BREIMAN, L. and FRIEDMAN, J. H. (1985), Estimating Optimal Transformations for Multiple Regression and Correlation, Journal of the American Statistical Association, 80,580-619.

COOK,R. D. and WEISBERG, S. (1982), Residuals and Influence in Regression, New York: Chapman and Hall.

LANDWEHR, J. M., PREGIBON, D., and SHOEMAKER, A. C., (1984), Graphical Methods for Assessing Logistic Regression Models, 79, 61-83

MALLOWS, C. L. (1982), Discussion of Atkinson, Journal of the Royal Statistical Society, 44, 29.

RAO, C. R. (1973), Linear Statistical Inference and It's Applications, John Wiley and Sons

RYAN, T., JOINER, B., and RYAN, B. (1976), Minitab Student Handbook, North Scituate, Mass. Duxbury Press.

Table of figure captions

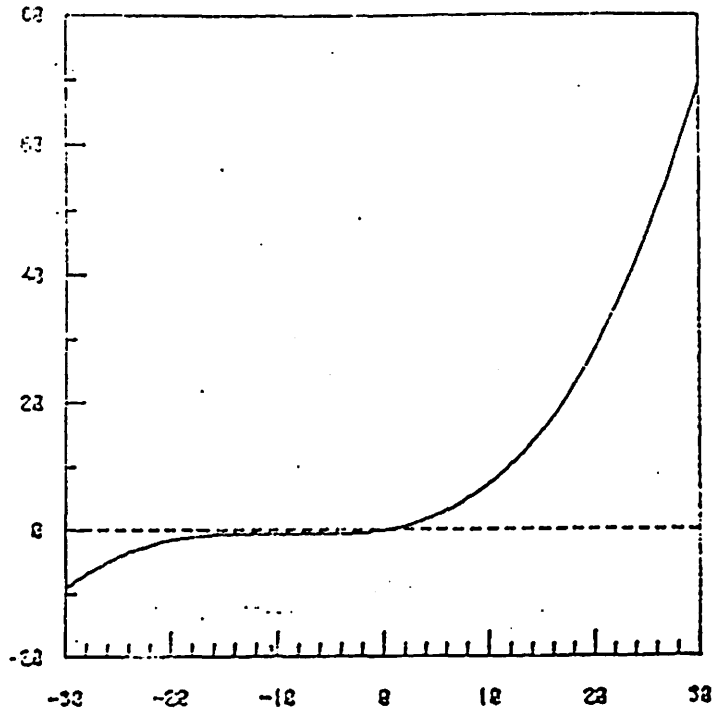

Figure 2.1

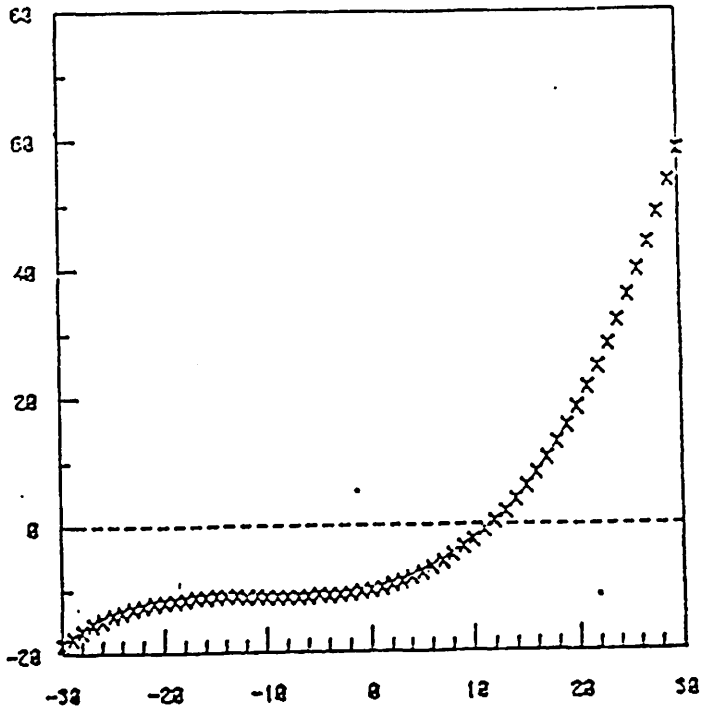

Figure 2.2

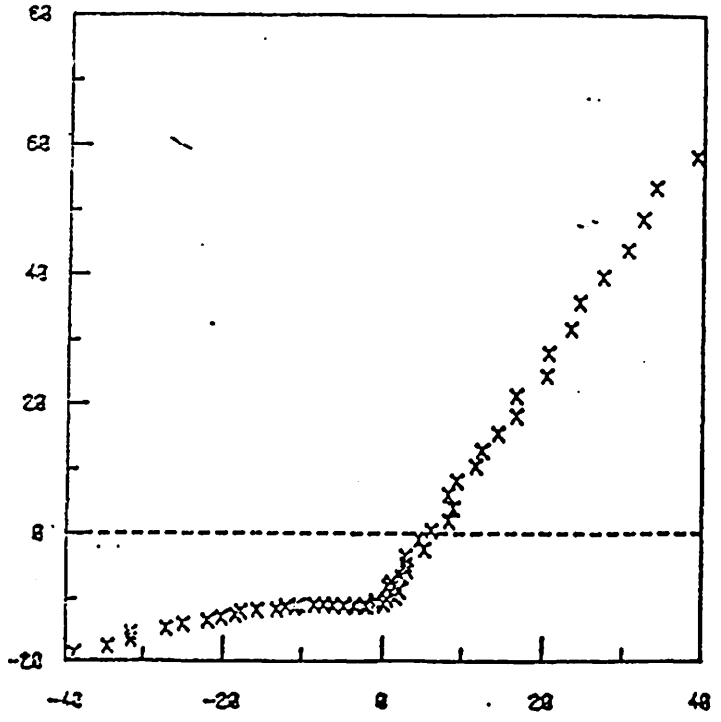

Figure 2.3

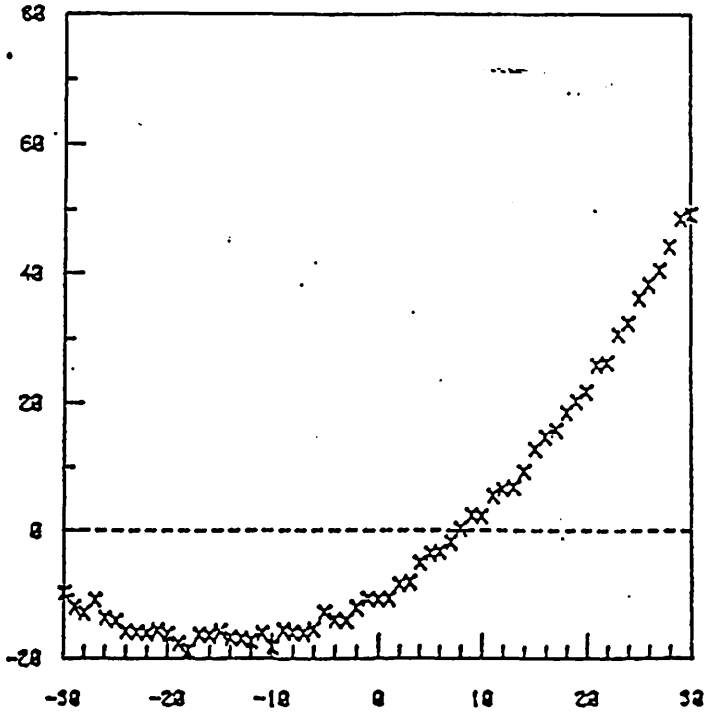

Figure 2.4

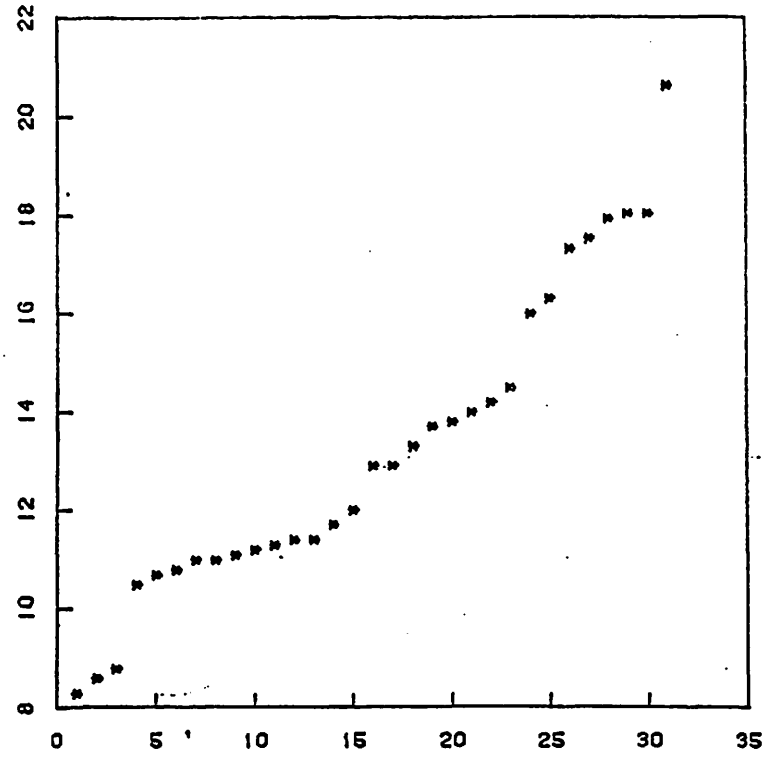

Figure 3.1.1

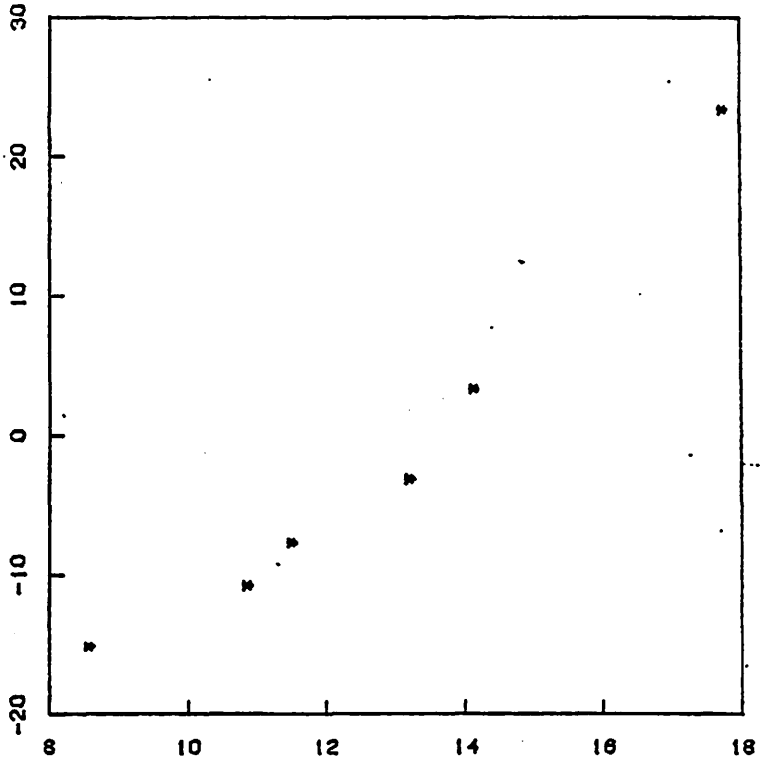

Figure 3.1.2

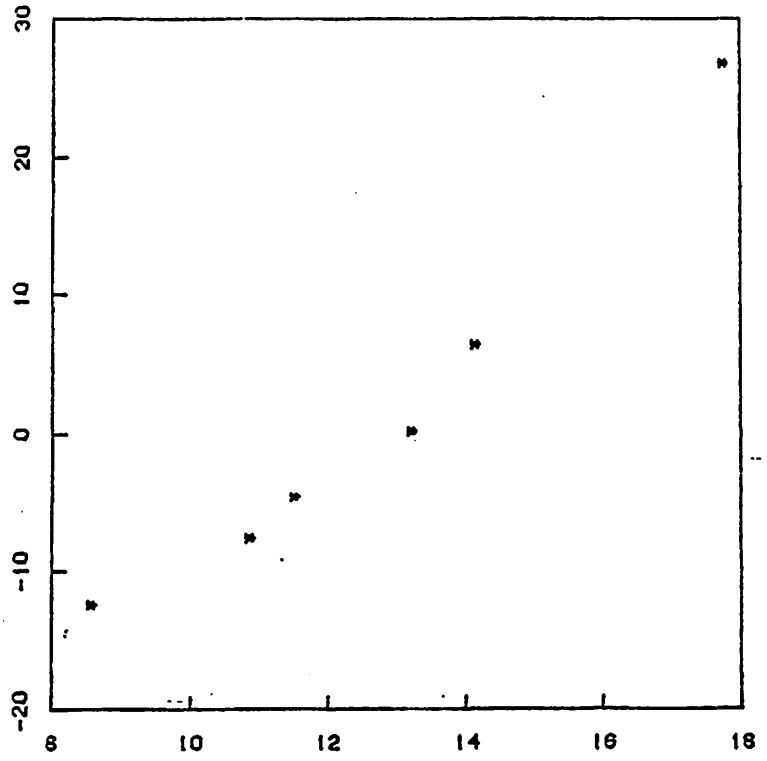

Figure 3.1.3

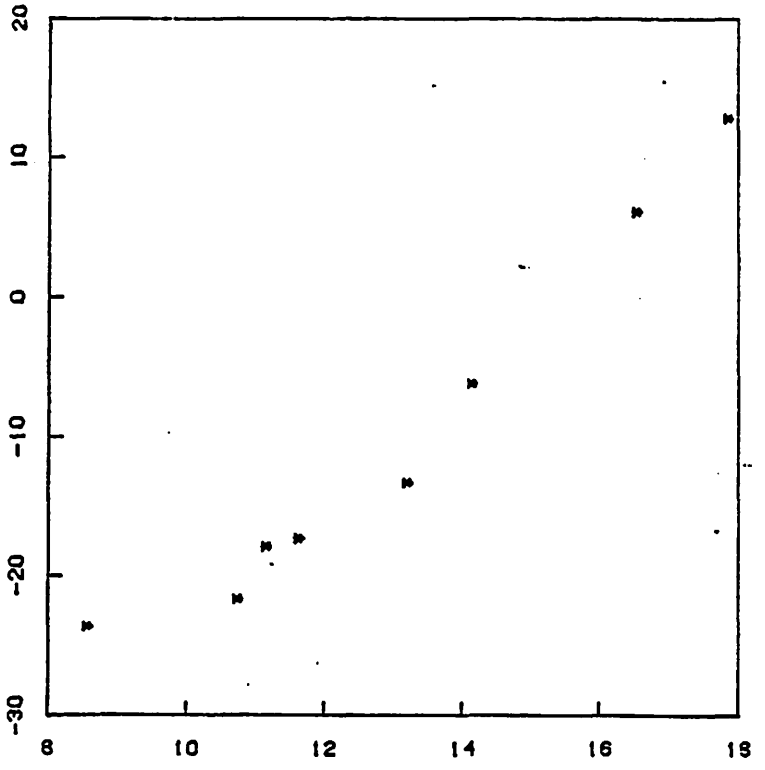

Figure 3.1.4

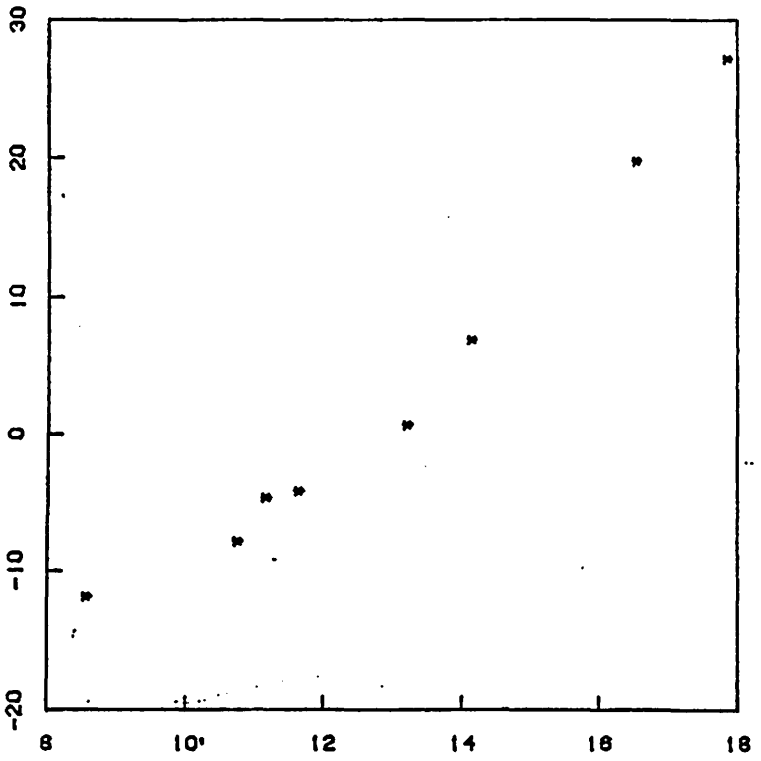

Figure 3.1.5

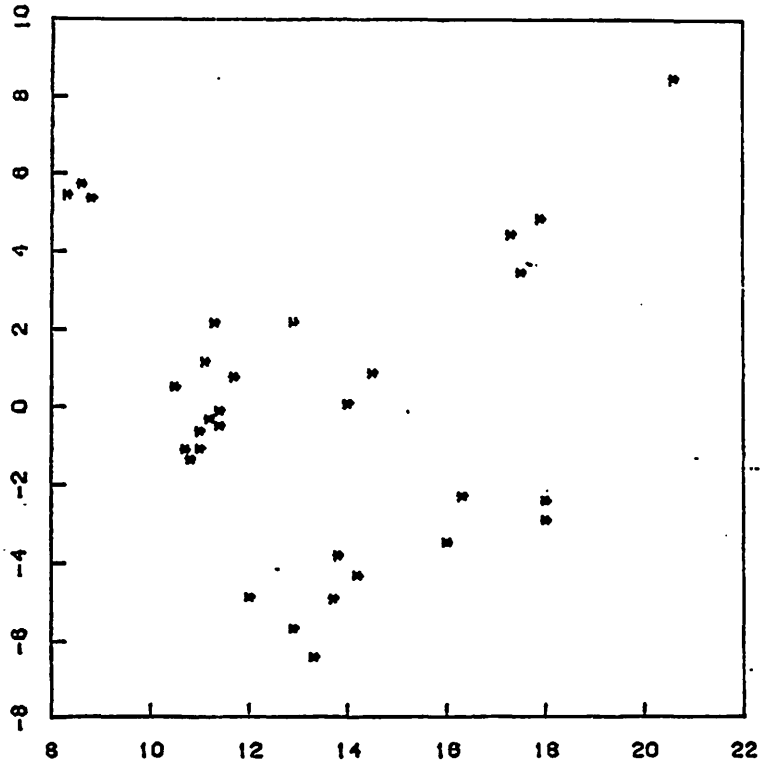

Figure 3.1.6

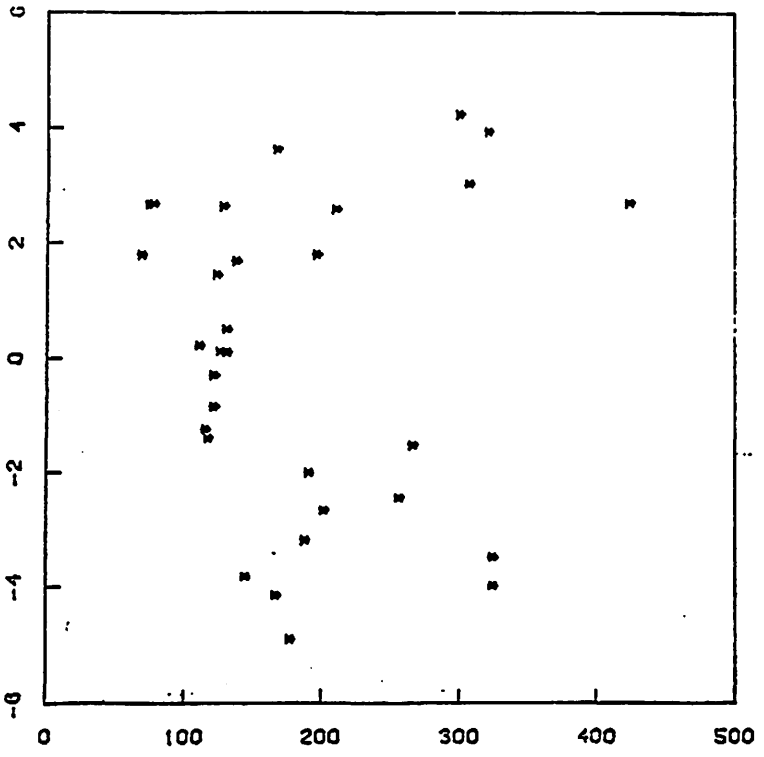

Figure 3.1.7

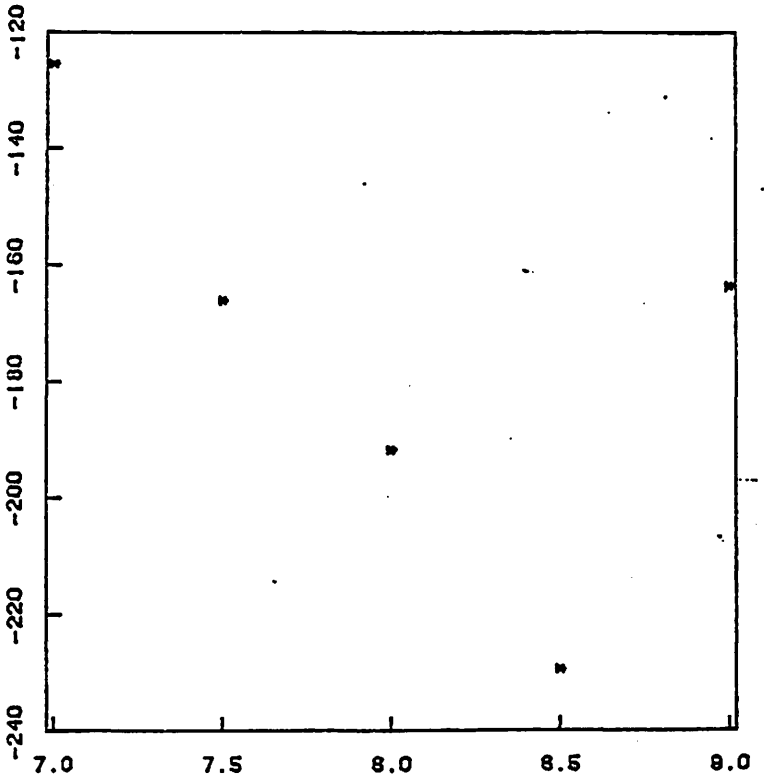

Figure 3.2.1

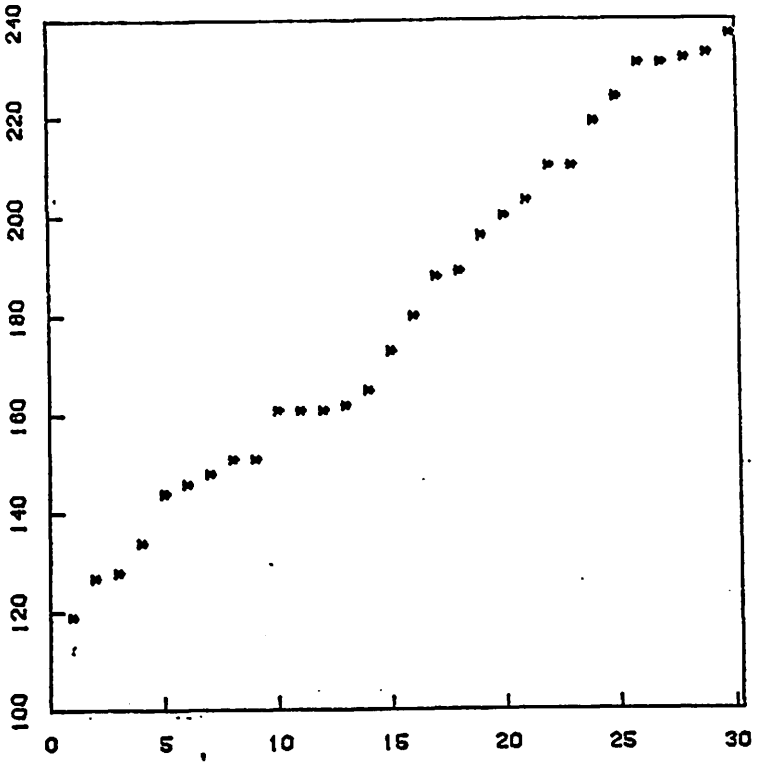

Figure 3.3.1

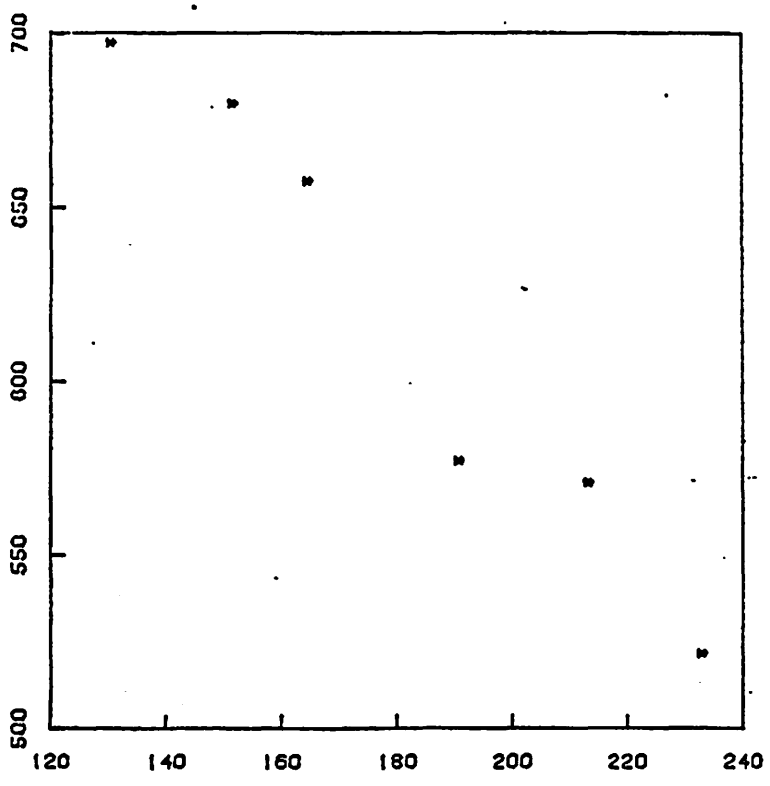

Figure 3.3.2

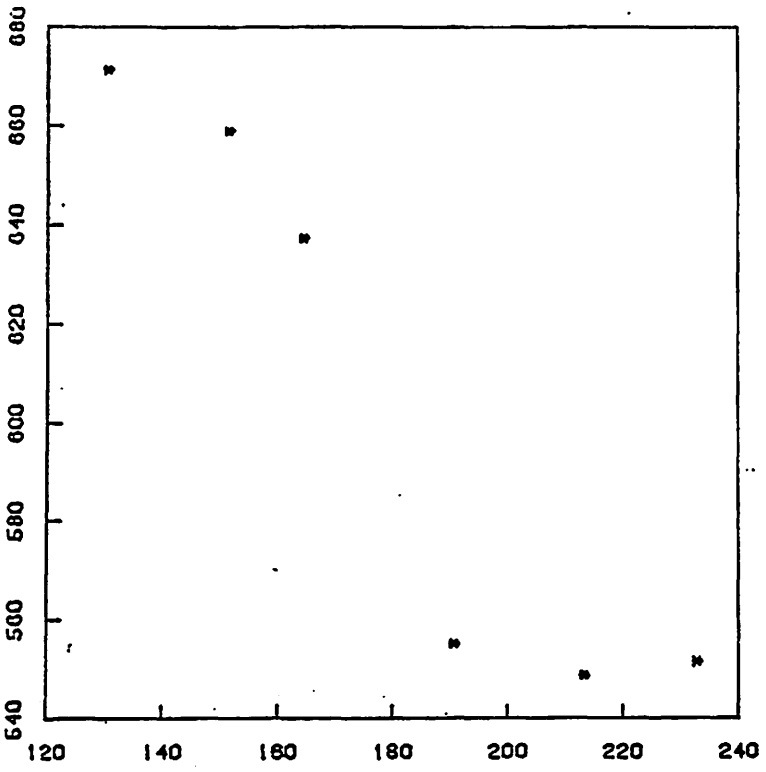

Figure 3.3.3

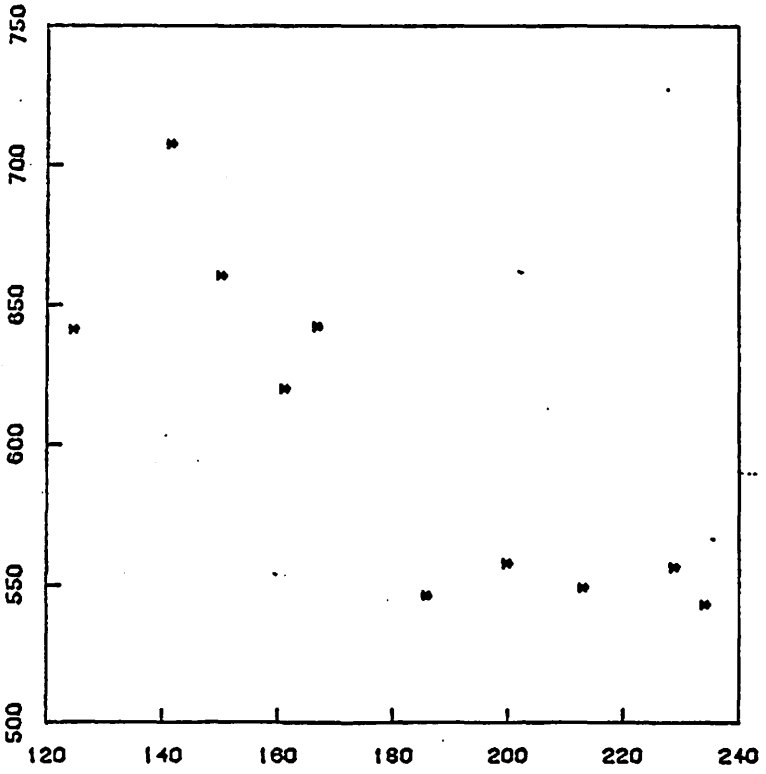

Figure 3.3.4

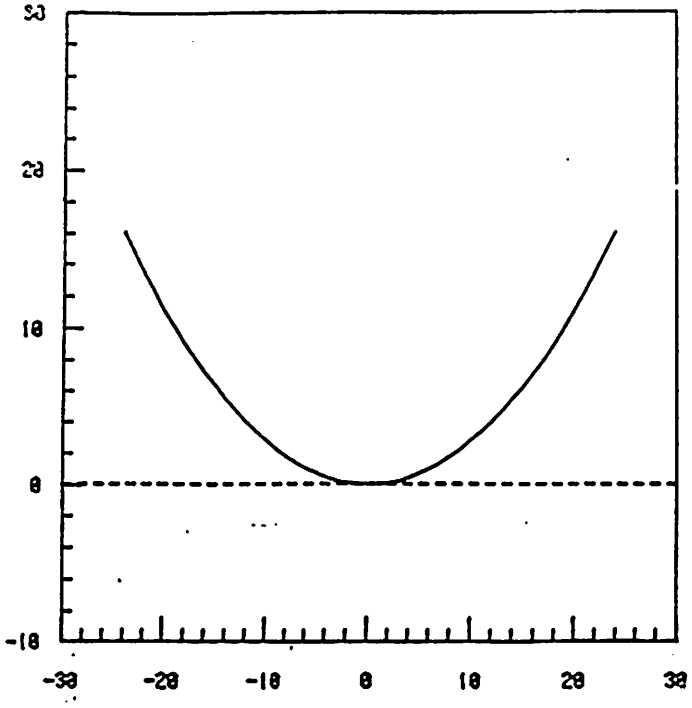

Figure 3.4.1

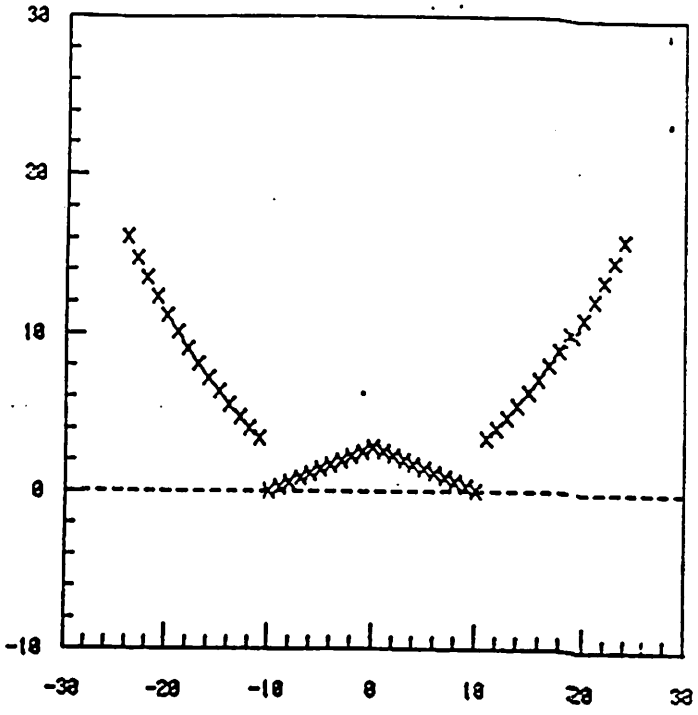

Figure 3.4.2

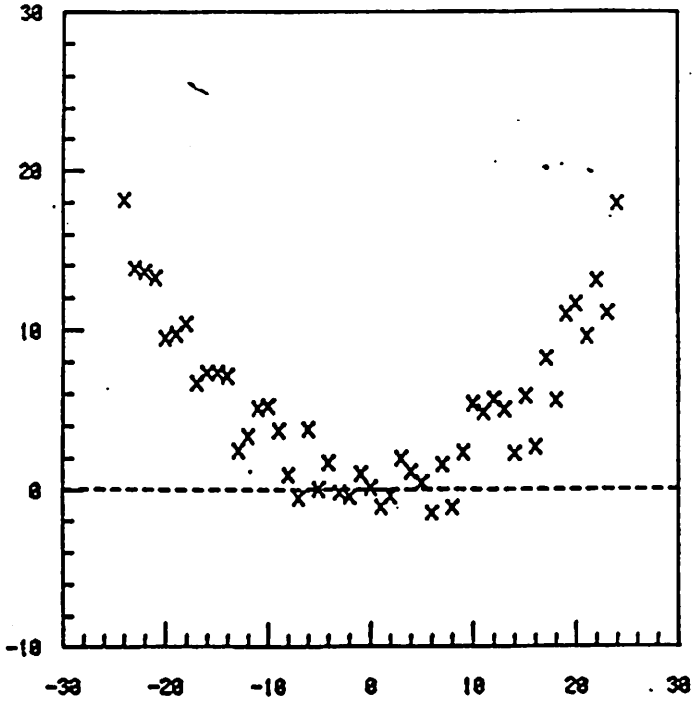

Figure 3.4.3

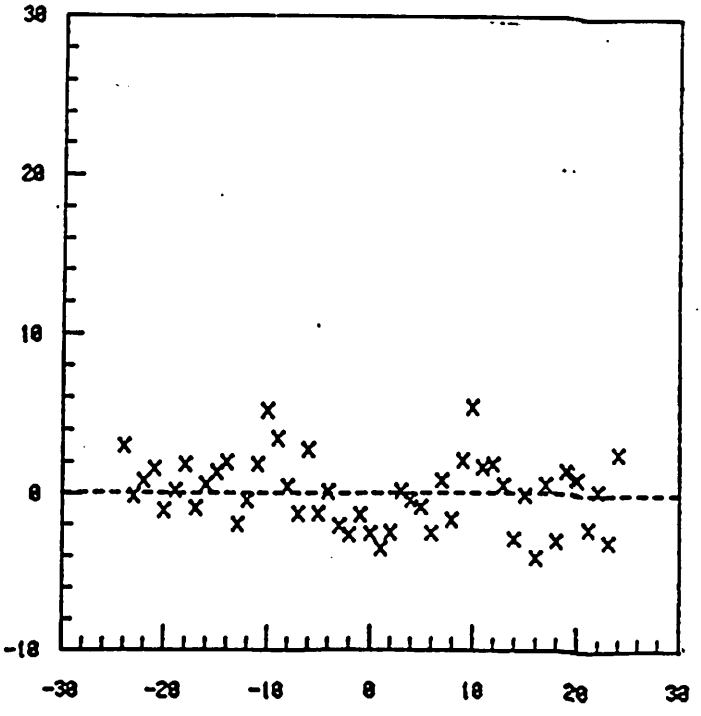

Figure 3.4.4

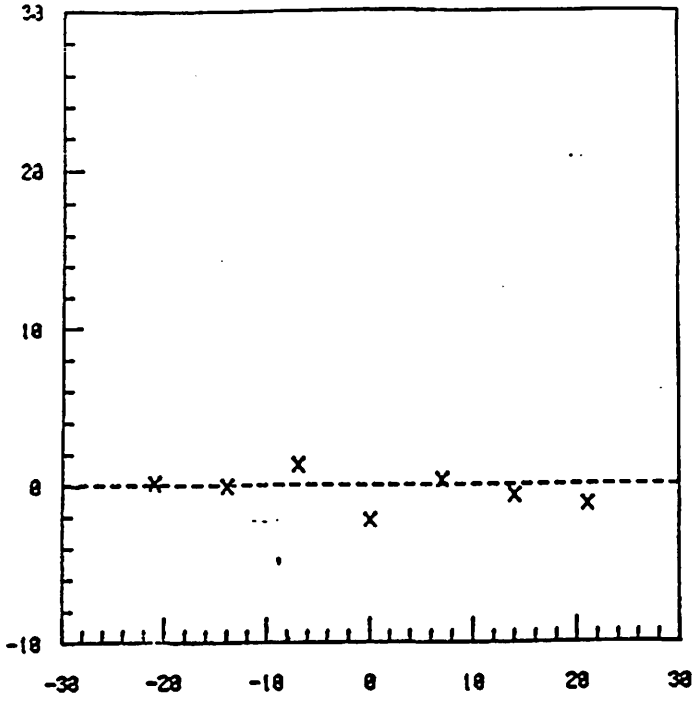

Figure 3.4.5

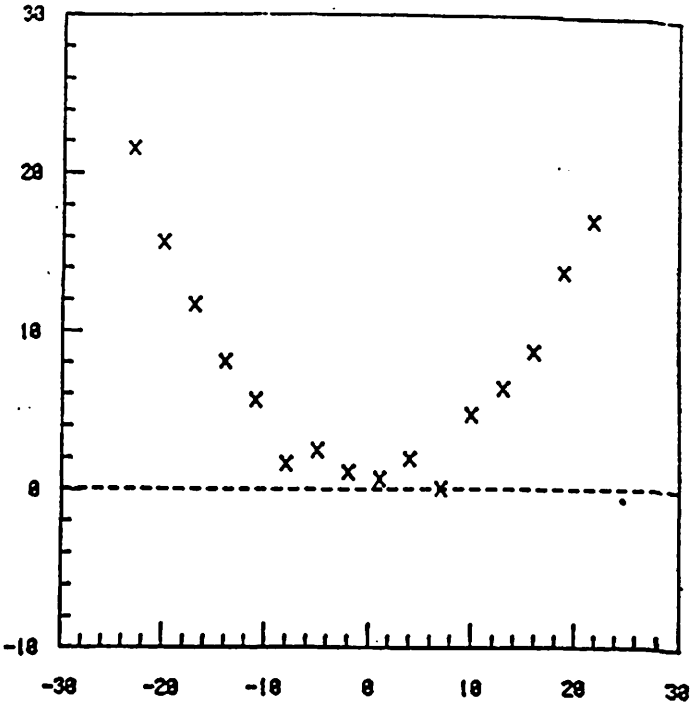

Figure 3.4.6

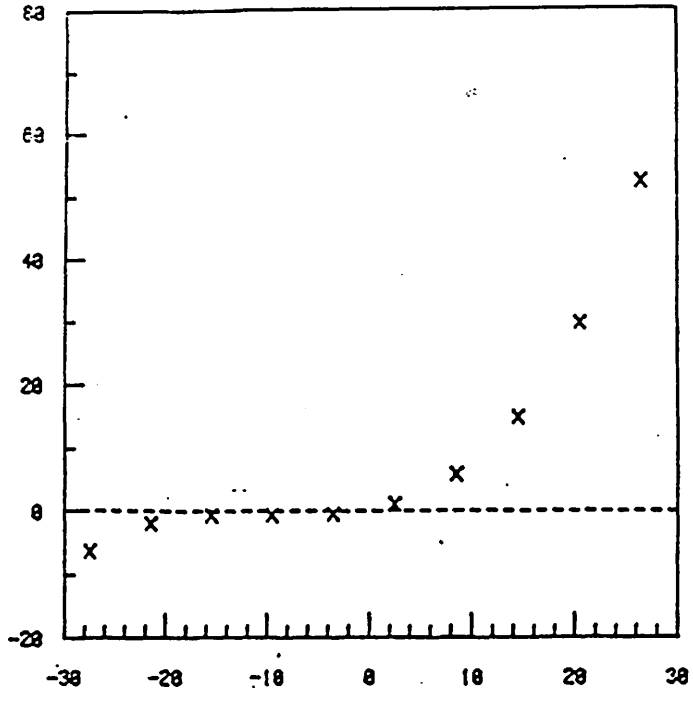

Figure 3.5.1

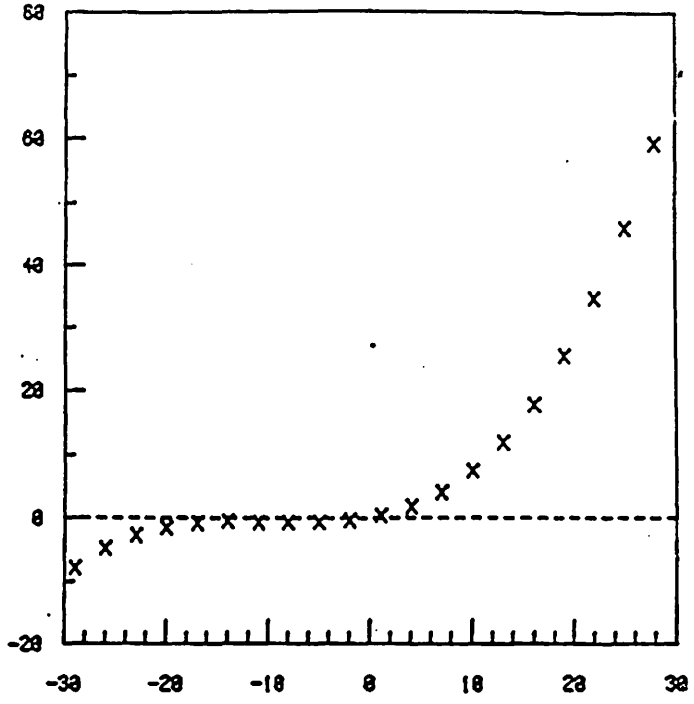

Figure 3.5.2